ASEAN Engineering Journal

Full Paper

AN ENSEMBLE APPROACH FOR COFFEE CROP YIELD PREDICTION BASED ON AGRONOMIC FACTORS

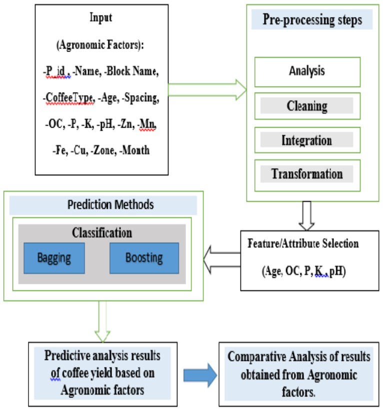
Article history
Received
08 July 2022
Received in revised form
29 January 2023
Accepted
05 February 2023
Published online
31 August 2023

Chandagalu Shivalingaiah Santhosh[a]*, Kattekyathanalli Kalegowda Umesh[b]*

[a]Department of Computer Applications, JSS Science and Technology, University, JSS Technical Institutions Campus, Mysuru, 570006, Karnataka, India.
[b]Department of Information Science and Engineering, JSS Science and Technology University, JSS Technical Institutions Campus, Mysuru, 570006, Karnataka, India.

*Corresponding author
Sancs84@jssstuniv.in

Graphical abstract


Input
(Agronomic Factors):
-P_id , -Name, -Block Name,
-CoffeeType, -Age, -Spacing,
-OC, -P, -K, -pH, -Zn, -Mn,
-Fe, -Cu, -Zone, -Month
Pre-processing steps
Analysis
Cleaning
Integration
Transformation
Feature/Attribute Selection
(Age, OC, P, K., pH)
Prediction Methods
Classification
Bagging
Boosting
Predictive analysis results of coffee yield based on Agronomic factors
Comparative Analysis of results obtained from Agronomic factors.

Abstract

Coffee is the most burned-through handled drink beside water, which is said to be the most exchanged cultivating product followed by oil in the entire globe. The two most significant sorts of coffee assortment filled in India are Arabica and Robusta out of 103 assortments of class coffee bean variety, which are economically exchanged around the planet. In this regard, we are taking major plantation crop in India i.e., Coffee for our research to explore and develop a predictive model for the development of coffee planters to take precise decisions in time during adverse situations in advance. Hence we propose a framework for coffee yield prediction which using machine learning ensemble approach to estimate the influence of agronomic factors to get a good coffee yield. Here, for our research work, the historic dataset is considered which is obtained from Central Coffee Research Institute (CCRI), Karnataka for the year (2008-2019). For the coffee yield prediction, we are considering agronomic factors like Age, Soil Nutrients: Organic carbon (OC), Phosphorus (P), Potassium (K), Alkaline (pH), Zone and Respective yield obtained in chikkamagaluru region, Karnataka state, India. Different classifiers are used namely, Extra Tree Classifier, Random Forest Classifier, Decision Tree and Boosting Algorithms for prediction and performance of each is compared and analyzed. Our results shown that Extra Tree Classifier and Random forest (RF) classifier with a precision of 91% with good results based on performance metrics considered respectively is an effective and versatile machine-learning method compared to other algorithms used.

Keywords: Machine Learning(ML), Ensemble, Coffee, Yield Prediction, Agronomic factors

1.0 INTRODUCTION

Coffee has been consumed by around 33\% of the world's population and projected that around 25 million grower families overall produce coffee, with a lion's share of smallholders and family members whose occupations mostly relies upon coffee as a ranch crop [1]. Coffee in India is more than a farmed export product, but it also a societal, institutional and traditional dry goods of southern states of India [2]. The two important types of coffee variety grown in India are C.Arabica and C.Robusta from out of 103 varieties of genus coffee which are commercially traded around the globe [3]. In Karnataka, Majority of the coffee planters are growing C.Arabica in Chikamagaluru and C.Robusta in Coorg regions respectively. However, there is decline in productivity and estimated that 10-20\% will be declined by 2050 due to agro-ecological factors in these regions [4]. In India, coffee producing is packed in the elevated plots of Southern states of India, among these states, 71\% of production accounted by Karnataka followed by 21\% of production from Kerala and 7\% of production from Tamil Nadu regions. Based on the survey, they have estimated around two fifty hundred thousand coffee bean producers in India of which 98\% of them are small-scale cultivators [5].

AI (ML) is applied based on the capacity of information pilot models to "learn" data with reference to a framework straightforwardly from noticed information by excluding foreordaining the orderly connections that deal with the framework.ML calculations can flexibly enhance their exhibition with every information test and find shrouded designs in complicated diversified and prominent measured information. AI has become the center innovation for various certifiable applications: from climate gauging and DNA sequencing, to web indexes and picture acknowledgment [6], [7], [8], and [9].

In this paper, we present a comprehensive literature review of the applications used for Coffee crop yield prediction based on agronomic factors. Andrew N Gillison et. al., have inspected the effect of various coffee trimming techniques on biodiversity (plant species richness) and soil nutrients for better coffee yield. Based on the data collected from different plots, they have done the comparison between soil texture and vegetation structure and cited that both the parameters relatively influence for better coffee yield [10].

Godsteven P Maro et. al., has developed a model for coffee yield estimation based on soil nutrient input device combined with root information collected from coffee growing areas of Tanzania. The outcome of this work is a novel framework, in which they thusly call it as Soil Analysis for Fertility Evaluation and Recommendation on Nutrient Application to Coffee. Later, this framework was examined for precision of the changed conditions, discovered to be equipped for imitating by 80-100% genuine yields [11].

Louis Kouadioa et. al., (2018) have assessed the capacity of a feed-forward network models called Extreme Learning Machine to break down clay richness belongings, also they have created an exact assessment of coffee yield. The exhibition of eighteen diverse Extreme learning machine-based models with available and different blends of the indicator factors dependent on organic matter components in soil was assessed. Toward the end, Extreme Learning Machine model's exhibition had been contrasted with that of current prescient devices like Multiple Linear Regression and Random Forest [12].

Y Romero-Alvarado et. at., (2002) have presented their study by evaluating the coffee yields based on soil nutrient contents. In this paper they have collected the data, which includes coffee production, soil-nutrient contents (Ca, Mg, K and P), soil-organic matter (SOM) and pH. At the end researchers have summarized the percentage of soil nutrients contributions are more for the consistent coffee yield [13].

Georg Rub et. al., (2008) have examined with suitable modeling techniques for wheat yield data. In this paper, they use feed-forward back propagation neural networks with seven parameters and evaluated based on the estimated results obtained. Also, economic and environmental parameters optimization of fertilization were carried out [16].

N Wanga et. al., (2014) have proposed regression analysis to study on yields and diverse production factors. The results of proposed algorithm, shown that biotic constraints and poor management practices were impacted on C.Robusta coffee yield production in the central region [17].

Christian Bunn et. al., (2015) have demonstrated an even minded way to deal with portray current and future environmental agro climatic changes spatially. They used random forest classification algorithm to display the spatial circulation of agro-natural zones. In addition to that, they have identified the set of climatic characteristics, these characteristics are evaluated C. arabica germplasm notwithstanding environment changes and to upgrade the assortments and other agronomic measures by thinking about the areas. At the end, the authors have concluded test of methodologies to improve assortments and other agronomic measures [18].

K Aditya Shastry et. al., (2016) have developed Custom-built ANN (C-ANN) for wheat crop yield forecasting by shifting the quantity of covered up hidden layers, number of neurons in the secret layer and the learning rate. Examinations were led to contrast of the algorithms used here on a similar informational collections Using R^2 measurement and rate expectation error. Outcome shown that, the C-ANN model achieved (97% test accuracy) with good R^2 value and model achieved lesser rate expectation error compared to Multiple Linear Regression with the accuracy of (92.52%) and D-CNN achieved (95% test accuracy) [19].

Sahu et. al., (2017) proposed a hadoop framework using random forest approach to foresee reasonable harvest for the field by thinking about different boundaries from soil and climate to examine the yields. In this paper, they have achieved 91% accuracy [20].

Subhadra Mishra et. al., (2016) directed an overview on different AI methods like neural networks, Random forest, greedy algorithms, Decision Tree, Regression methods, Time series models and different clustering methods were adapted on agricultural crop production sector for better yield predictions for the crops like corn, cotton, rice, sugar, sugarcane, wheat, soyabean and jowar respectively [21].

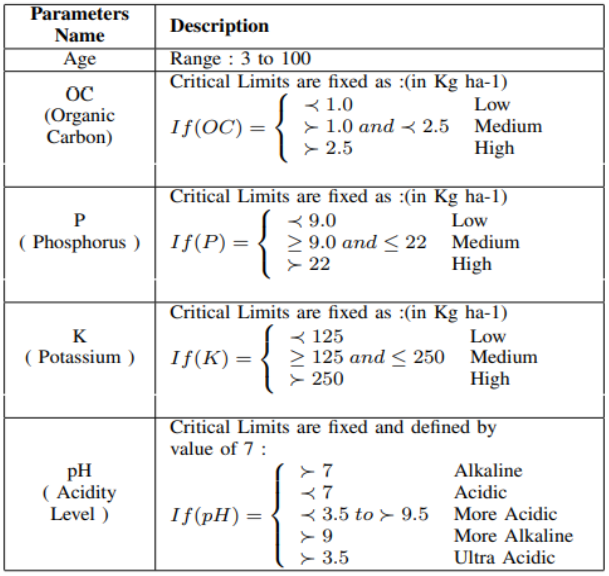
To develop crop yield prediction model by considering agronomic factors for chikkamgaluru district, we carried out a literature survey on various machine learning techniques on agricultural crop production domains. In this unique circumstance, we found that, numerous scientists have proposed information mining and AI strategies to create predictive models. In section-2, we proposed and discussed machine learning models for the forecast of coffee yield based on agronomic factors. In this section, we discussed about data sets and methods in detail. In section-3, the results are tabulated and presented in detail. In section-4, we discussed three different models were used to predict coffee yield in detail.

2.0 DATASET AND METHODS

2.1 Study Area and Dataset

The dataset are collected from Central Coffee Research Institute (CCRI), Balehonnur, Karnataka for the year 2008-2019 of Chikkamagaluru district, Karnataka State, India. The dataset contains agronomic factors, comprised with seven input features of 19898 samples of nine zones labeled from 1 to 9 range for the model of our research work. The portrayal of the datasets are illustrated in the Tables 1 and 2 respectively. Table 1 describes Five Continuous parameters and Table 2 describes Two Categorical parameters in detail.

Table 1 Five Continious Parameters



Parameters Name	Description
Age	Range : 3 to 100
OC (Organic Carbon)	Critical Limits are fixed as :(in Kg ha-1) $If(OC) = \begin{cases} \prec 1.0 & \text{Low} \\ \succ 1.0 \; and \prec 2.5 & \text{Medium} \\ \succ 2.5 & \text{High} \end{cases}$
P (Phosphorus)	Critical Limits are fixed as :(in Kg ha-1) $If(P) = \begin{cases} \prec 9.0 & \text{Low} \\ \geq 9.0 \; and \leq 22 & \text{Medium} \\ \succ 22 & \text{High} \end{cases}$
K (Potassium)	Critical Limits are fixed as :(in Kg ha-1) $If(K) = \begin{cases} \prec 125 & \text{Low} \\ \geq 125 \; and \leq 250 & \text{Medium} \\ \succ 250 & \text{High} \end{cases}$
pH (Acidity Level)	Critical Limits are fixed and defined by value of 7 : $If(pH) = \begin{cases} \succ 7 & \text{Alkaline} \\ \prec 7 & \text{Acidic} \\ \prec 3.5 \; to \succ 9.5 & \text{More Acidic} \\ \succ 9 & \text{More Alkaline} \\ \succ 3.5 & \text{Ultra Acidic} \end{cases}$

Table 2 Two Categorical Parameters

Parameters Name	Description
ZONE	The Zones are labelled with numbers as: Balehonnur-1, Belur-2, C Chikkamagaluru Town-3, Hassan-4, Kalasa-5, Koppa-6 Mudigere-7, N R Pura-8 and Sakaleshpura-9.
Yield	Yield as analysed and labelled between Class-1 to Class-4 as follows: - O to 300kg : Class-1 - 301kg to 600Kg : Class-2 - 601Kg to 900Kg : Class-3 - 901Kg to 1200Kg : Class-4

2.2 Proposed Methodology

In this segment, we put forward machine learning models to develop and predict the coffee yield based on agronomic factors. The square outline of proposed methodology is illustrated in Figure 1.

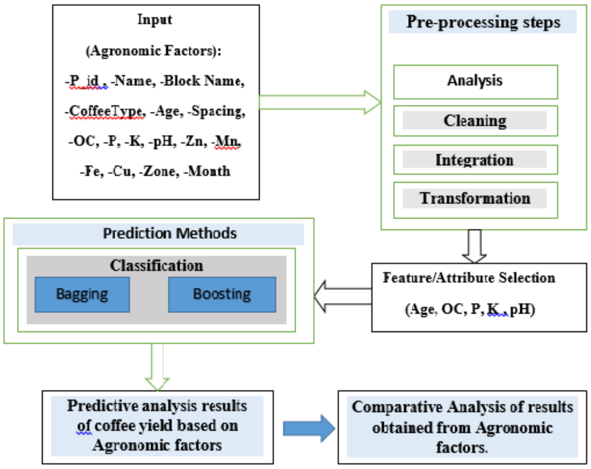

Input
(Agronomic Factors):
-P_id , -Name, -Block Name,
-CoffeeType, -Age, -Spacing,
-OC, -P, -K, -pH, -Zn, -Mn,
-Fe, -Cu, -Zone, -Month
Pre-processing steps
Analysis
Cleaning
Integration
Transformation
Feature/Attribute Selection
(Age, OC, P, K , pH)
Prediction Methods
Classification
Bagging
Boosting
Predictive analysis results of coffee yield based on Agronomic factors
Comparative Analysis of results obtained from Agronomic factors.

Figure 1 Proposed Methodology block diagram for Coffee Yield Prediction System

In this proposed work, we have scrutinised the soil sample data which was provided in the format of historical (Numerical) data in the form of ledgers by central coffee research institute (CCRI), Balehonnur, Karnataka laboratory. This historical data from 2008 to 2021 was further analysed and manually transformed into CSV data format. In the beginning, we have verified the data sets and understood that few missing values were present. These missing values are filled by calculating statistical approaches like mean, median and binning methods. Based on the domain experts inputs from CCRI laboratory, the five features/Attributes has been chosen empirically which describes/influences towards good coffee crop yield, which is mentioned in Table-1.Then we applied boxplot on chosen soil factors to remove the outliers based on the standard critical limits that is formed by CCRI on OC, P, K and pH [23]. After eliminating the outliers the data set is reduced from 20,073 to 15,755 instances and then we applied skewness measure to verify the distribution of the data set [22].

2.3 Methods

In this section, we discussed about a machine learning ensemble approaches.

Decision Tree Classifier (DTC): Decision Tree Classifier (DTC) algorithm is popular algorithm for classification and prediction [15]. This Algorithm is used to develop a prediction model in terms of predictive variables. The Decision Tree Classifier algorithm is illustrated in the following steps:

```
AB(C, Aspect, Target)
 n = create Node ()
 Label (n) = most Common Class (C, Target)
 IF v hx, d (j)i € C:d(j) = d  THEN  return(n)  ENDIF
 IF Aspect = Ø THEN   return (n) ENDIF
 R*= argmaxR € Aspect (information Gain(C, R))
 For Each   r € R * DO
      Cr = {(x, d (j)) € C: j |R| = r}
      IF Cr = Ø THEN
          nO =  create  Node()
```

```
        Label (nO) = most Common Class (C, Target)
        Create Edge (n, r, nO)
     ELSE
         Create Edge (n, r, AB (Cr, Aspect R*, Target))
     END IF
END DO
return (n)
```

The tree is constructed as indicated by the split criterion until to reach the ideal decision tree hierarchy nodes. The Gini Diversity Index (GDI) is considered as split optimization criterion, to quantify the node impurity. The measure of Gini Index, at a node D is calculated using equation number-1.

$$Gini = 1 - \sum_{i=1}^{c} (\pi)^2 \quad \text{-------- (1)}$$

Random Forest Classifier (RFC): The idea of using Random Forest (RF) Classifier algorithm is to build a small classification Tree (DT) in parallel with minimal number of features, at that point consolidate the trees to frame a unique robust learner by meaning the greater number of votes. It often found to be the accurate learning algorithm and computationally cheap process. It is used to build an ensemble of classification tree usually trained with bagging method [12].

The Random forest (RF) classifier algorithm is illustrated in the following steps:

Training: A Preparation set D: = (e1, f1)… (en, fn), features Z, also the number of trees in forest F.

Function Random Forest (D, Z)

```
   W ← Ø
   for i € 1,....,F do
        D (i) ← A   bootstrap sample from D
                wi← Randomized  Tree  Learn (D(i), Z)
                W ←W U {wi}
   end for
   return W
end function
```

Function Randomized Tree Learn (D, Z)

```
 At each node:
    z ← very  small  subset  of   Z
                  Split on best feature in   z
             return The  learned $tree
 end function.
```

Extra Trees Classifier (ETC): Extra Trees Classifier is an outfit AI calculation that consolidates the expectations from numerous choice trees. This method is utilized to construct the decision tree as members of the ensemble [14].

The Extra tree classifier algorithm is illustrated in the following steps:

Split a node (D)

```
Input: the nearby learning subset D relating to the node we need to split
Output: a node split [m ¦ mn] or zero split
If Stop split (D) == then return zero.
In any case select P attributes {b1... bp} among all non-Consistent (in D) applicant attributes;
 Illustrate P splits d1,...,dp , whereabouts di = choose a arbitrary
              Split (D, mi), v j = 1... P;
Return a split d * such that   Count (d*, D) = maxi=1... P Count (di, D).
```

Pick a random split} (D,m)

```
Inputs: m subclass D and an attribute m
Output: m split
Let mDmax and mDmin indicate the maximal and negligible estimation of m in D;
 illustarte a arbitrary border ac correspondingly in [mDmin, mDmax]
 Return the split [m < mn].
```

```
     Stop split (D)
     Input: m subclass D
    Output: m binary
    If |D| <  xmin , then
         return TRUE;
  In the event that all attributes are consistent in D, then
  return TRUE;
  If the outcome is consistent in D, then
  return TRUE;
if not,  return FALSE.
```

Here the decision criteria used will be Information Gain. First, we calculate the entropy of the data and then the information gain, with the following equation number-2:

Entropy(S) = - C $\sum i\pi log_2\pi$ -------- (2)

Information Gain = entropy (parent) – [average entropy (children)].

Boosting Algorithms: Boosting algorithms are a set of helping calculators and are a bunch of the low precise classifier to make a profoundly exact classifiers. Low precision classifier (or frail classifier) offers the exactness better than the flipping of a coin. Exceptionally precise classifier (or solid classifier) offer mistake rate near 0.Boosting algorithm can follow the model who bombed the exact forecast. Helping calculations are less impacted by the over fitting issues on the models.

a. Gradient Boosting: In gradient boosting ensembles are added in stages. In every stage, weak leaners are added to compensate for the existing weak learners. The shortcomings of the combined model are identified with gradients. In our study, it reduced the variance and bias because of its sequential classifiers [24].

The Gradient boosting algorithm is illustrated in the following steps:

Step 1: start with primary guess

The underlying theory of the Gradient Boosting calculation is to foresee the typical worth of the objective y. For instance, in the event that our elements are the age x1x1 and the level x2x2 of an individual and we need to anticipate the heaviness of the individual.

Step 2: Process the pseudo-residuals
For the variable x1, we register the contrast between the perceptions and the expectation we made. This is known as the pseudo-residuals.
Step 3: Predict the pseudo-residuals
Then, we will be using the features x1, x2, x3, x4 to predict the pseudo-residuals column.
Step 4: Make an expectation and process the residuals
Step 5: Make a second prediction
Presently, we:
-construct a subsequent tree
-register the forecast utilizing this subsequent tree
-process the residuals as per the expectation
-fabricate the third tree
etc

b. XG Boosting : In this algorithm, decision trees are made in consecutive structure. Here loads assume a significant part in XGBoost. Loads are allocated to every one of the autonomous factors which are then taken care of into the choice trees which predicts the outcomes. Likewise loads of factors anticipated wrong by the tree is expanded and these factors are then taken care of to the subsequent choice tree. Consequently individual classfiers/indicators then outfit to give serious areas of strength for a more exact model [25].

In our study, we have built a tree of 100 with max depth of 7 of the classification problems by specifying the binary attribute with decision tree, random forest and extra tree classifiers.

The XG boosting algorithm is illustrated in the following steps:
Step 1 – Creating the First Base Learner
Step 2 – Calculating the Total Error (TE)
The total error is the amount of the multitude of blunders in the arranged record for test loads. For our situation, there is just 1 mistake, so Total Error (TE) = 1/5.
Step 3 – Computing the performance of the Stump.

$$\textbf{Performance of Stump} = 1/2 \ln \frac{[\,1\text{-}TE\,]}{TE} \quad \text{---------}\ (3)$$

where, ln is natural log and TE is Total Error.
Step 4 – Updating Weights
Step 5 – Creating a New Dataset
Step 6 - Follow from step 3 to step 5

c. **Ada Boosting**: In this algorithm, different classifiers are consolidated to expand the precision of classifiers. AdaBoost is an iterative troupe strategy which fabricates serious areas of strength for a by joining various ineffectively performing classifiers with the goal that you will get high areas of strength for precision. The fundamental idea driving AdaBoost is to set the loads of classifiers and preparing the information test in every emphasis to such an extent that it guarantees the precise expectations of strange perceptions. Here, any AI calculations can be utilized as base classifier assuming it acknowledges loads on the preparation set [26].

In our study, 100 weak learners are set to train iteratively and improved the model for the base_estimators like Decision tree, Random Forest and Extra Tree Classifiers algorithms.

The Ada boosting algorithm is illustrated in the following steps:

Step 1: Allocate Equal Weights to every one of the perceptions.
At first allot same loads to each record and keep in the dataset
Test weight = 1/N [Where N = Number of records]
Step 2: Characterize arbitrary examples utilizing stumps
Draw arbitrary examples with supplanting from unique information with the probabilities equivalent to the example loads and fit the model.
Step 3: Compute Total Error
Complete mistake is only the amount of loads of misclassified record. Total Error = Weights of misclassified records.
Complete mistake will be somewhere in the range of 0 and 1 ,100% of the time. 0 addresses amazing stump (right arrangement) and 1 addresses weak stump (misclassification).
Step 4: Compute Performance of the Stump
Utilizing the Total Error, decide the Execution of the base student. The Determined execution of stump(α) Esteem is used to refresh the loads in Sequential emphasis and furthermore
Utilized for conclusive forecast computation.
Step 5: Update Weights
In light of the exhibition of the stump(α)
Update the loads. We really want the following stump to accurately arrange the misclassified record by expanding the comparing test weight and diminishing the example loads of the accurately arranged records.
Step 6: Update weights in iteration
Utilize the standardized weight and make the second stump in the woods. Make a new dataset of same size of the first dataset with reiteration in view of the recently refreshed test weight.
So that the misclassified records get higher likelihood of getting chosen. Rehash Step 2to 5 again by refreshing the loads for a
Specific number of emphases.
Step 7: Final Predictions
Last forecast is finished by getting the indication of the weighted amount of last anticipated esteem.

Model Performance Evaluation

Predictions made by the calculations utilized are constant genuine esteemed numbers in the reach between 1 to 4, which describe the probability of yield. To change over the nonstop forecasts into class names 50% cutoff tip was enforced. The contrast among anticipated and anticipated class results was then portrayed by various True Positive's (TP), True Negative's (TN), False Positive's (FP) and False negative's (FN), were the amount of TP+TN+FP+FN = n is the complete number of perceptions.

The Table 3 gives the Standard Performance Measurements, which were utilized to survey the precision of the calculations utilized:

Table 3 Standard Performance Metrics

Sl No	erformance Metrics	Formula
1.	Correct Classification Rate [C]	C = (TP + TN) / n
2.	Precision	Precision = TP/(TP+ FP)
3.	Recall	Recall = TP / (TP + FN)
4.	F1 – Score	F1 = 2 * (Precision * Recall) / (Precision + Recall)
5.	Receiver Operating Characteristic (ROC) Curve	As a rule of thumb, we will assess the performance of the model as : 0.9 to 1.0 = A (Very Good) 0.8 to 0.9 = B (Acceptable) 0.7 to 0.8 = C (Impartial) 0.6 to 0.7 = D (Unsatisfactory) 0.5 to 0.6 = E (reject)

3.0 RESULTS

Random Forest Classifier (RFC): A Random Forest of 100 tree accomplished C = 91% throughout the training stage also effectively grouped with same 91% of experiments, which is closely resembling the Random Forest model execution on a similar test associate introduced in (Table 4). At the point when further assessed on the test associate, Random Forest performed with 91% Recall and the AUC test of 0.960 for 50:50 split ratio (Figure 2).

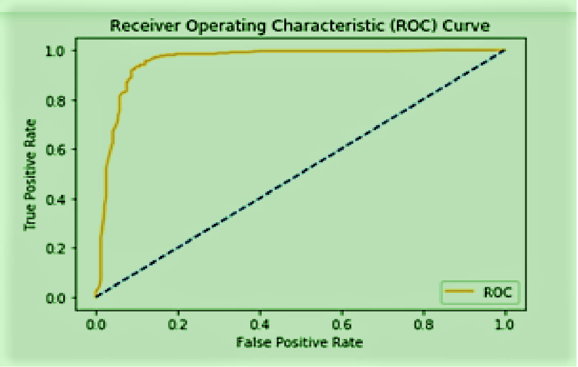


Figure 2 Receiver Operating Characteristic Curve for Random Forest Model precision on the test samples obtained for 50:50 split ratio.

Experiment with RF was repeated with 10 Kfold to decide if the consistency improved contrasted with the DT model. The outcomes showed essentially same and overall, consistently performance for different split is shown in Table 4. The Key important features considered for RF are the 4 soil parameters OC, P, K, pH and Age used by RF classifier which will result in identifying the Yield class.

Table 4 Predictive Performance of RFC Model

Sl No	Split Ratio	Accuracy	Precision	Recall	F1-Score	ROC Curve
1	70:30	90.0	91.0	90.0	90.0	0.96
2	60:40	92.0	92.0	92.0	92.0	0.96
3	50:50	91.0	91.0	91.0	91.0	0.96

The confusion matrix for testing data is shown in figure 3. The algorithm identifies 215 genuine positive rate and 242 genuine negative rates for training data, whereas for the test date 215 genuine positive rate and 242 genuine negative rates. 17 of the observation are negative, but it predicted positive whereas 26 observation are positive and it predicted negative for training data. 17 of the observation are negative, but it predicted positive whereas 26 observation are positive and it predicted negative for testing data.

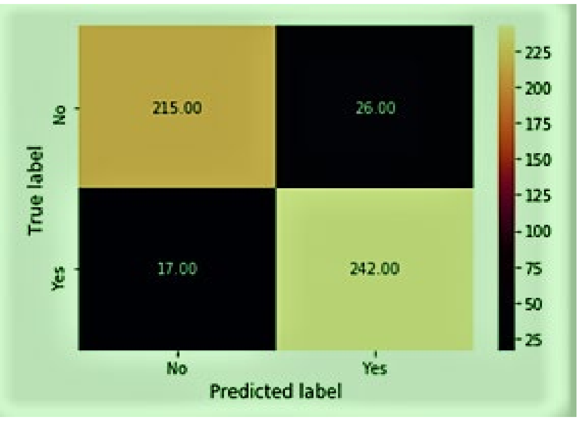


Figure 3 Confusion Matrix for the test samples obtained on 50:50 split ratio for RF model

Extra Tree Classifier (ETC): Like a Random Forest classifier we have the Extra Trees classifier — otherwise called Extremely Randomized Trees. All the data accessible in the preparation set is utilized to assemble every node. To shape the root node or any other node, the best split is controlled via looking in a subset of arbitrarily chose features of size sqrt (number of features). The split of each chose feature is picked aimlessly.

The ETC model was created subsequent to considering 100 DTs based on various subgroups of the information by modifying the test and train datasets with one another. The ETC had the option to accurately predict the Yield in C = 91% when three different split cases are considered on test datasets (Table 5).

Table 5 Predictive Performance of ETC Model

Sl No	Split Ratio	Accuracy	Precision	Recall	F1-Score	ROC Curve
1	70:30	90.0	92.0	90.0	91.0	0.95
2	60:40	92.0	92.0	91.0	91.0	0.95
3	50:50	91.0	91.0	91.0	91.0	0.95

The ETC identified yield for testing data with Recall of 91%, and AUC test of 0.95 for 50:50 split ratio (figure 4).

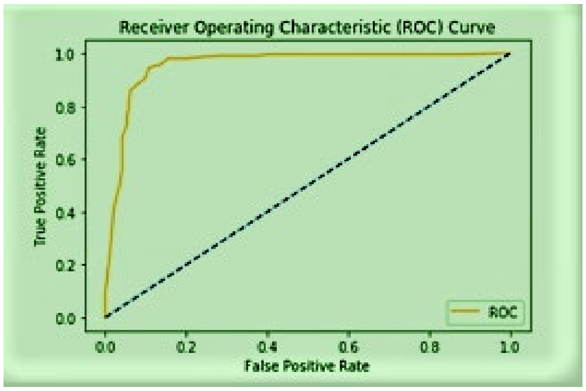


Figure 4: Receiver Operating Characteristic curve of ETC Model precision on the test samples obtained for 50:50 split ratio

The confusion matrix for testing data is shown in figure 5. The algorithm identifies 215 True positive rate and 245 True negative rates for the test data. 14 of the observation are negative, but it predicted positive whereas 26 observation are positive and it predicted negative for testing data.

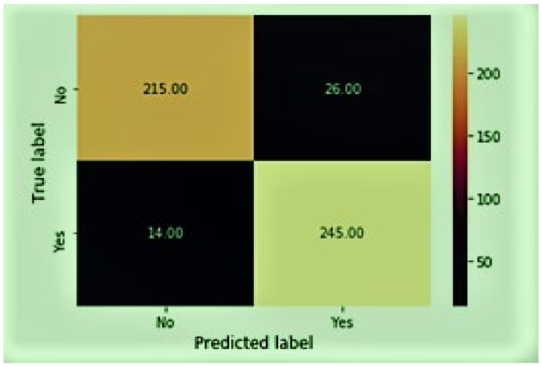


Figure 5 Confusion Matrix for the test samples obtained on 50:50 split ratio for ETC mode*l*

Decision Tree Classifier (DTC): The DT model was created subsequent to considering 100 DTs based on various subgroups of the information by modifying the test and train datasets with one another. The decision tree had the option to effectively predict occurrence of Yield in C = 84% when three different split cases are considered on preparing and test datasets (Table 6).

Table 6: Predictive Performance of DTC Model

Sl No	Split Ratio	Accuracy	Precision	Recall	F1-Score	ROC Curve
1	70:30	83.0	83.0	83.0	83.0	0.83
2	60:40	84.0	84.0	84.0	84.0	0.84
3	50:50	84.0	84.0	85.0	84.0	0.84

When we estimated on the test group, the Decision Tree recognized yield +ve with 84% Recall (Table 6). Figure 6 depicts the classifier Receiver Operating Characteristics curve with Area under ROC Curve test = 0.84 for the decision tree forecasts on tests for 50:50 Split ratio (Figure 6).

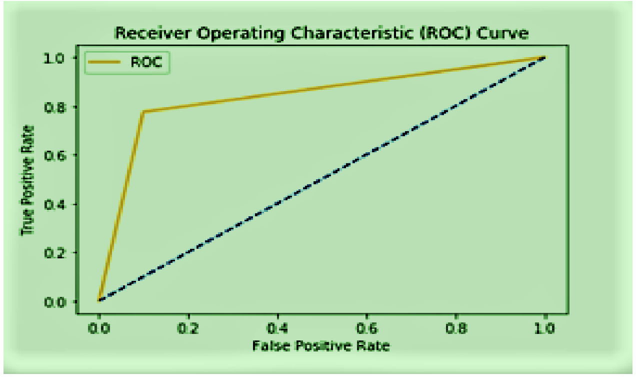


Figure 6: Receiver Operating Characteristic curve for DTC Model precision on test samples obtained for 50:50 split ratio

The confusion matrix for testing information is shown in figure 7. The algorithm identifies for the test date 217 genuine positive rate and 201 genuine negative rates. 58 of the observation are negative, but it predicted positive whereas 24 observation are positive and it predicted negative for testing data.

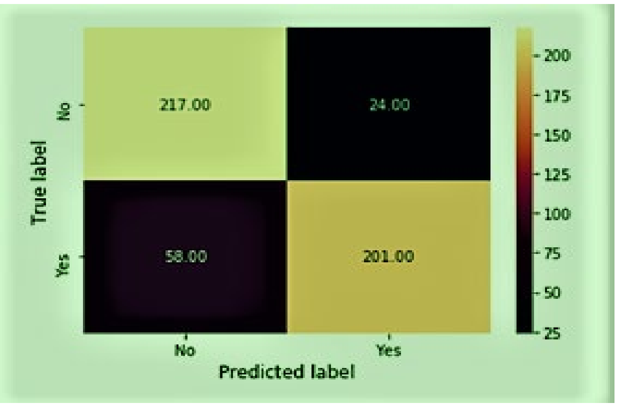


Figure 7 Confusion Matrix for the test samples obtained on 50:50 split ratio for DTC model

Boosting Algorithms:

Here the boosting technique is applied with 100 weak learners by considering the base_estimator as random forest. This technique is considered for the three different boosting techniques Gradient Boosting, XG Boosting and Ada Boosting.

a. Gradient Boosting:

Here the boosting is applied with 100 weak learners by considering the base estimator as random forest. This technique is considered boosting technique: Gradient Boosting Algorithm.

Table-7 Predictive Performance of Gradient Boosting Model.

Sl No	Split Ratio	Accuracy	Precision	Recall	F1-Score	ROC Curve
1	70:30	91.0	93.0	88.0	90.0	0.95
2	60:40	92.0	94.0	89.0	91.0	0.95
3	50:50	90.0	91.0	90.0	90.0	0.95

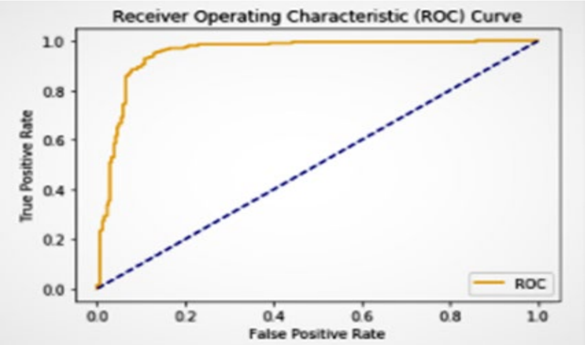


Figure 8. Receiver Operating Characteristic curve for Gradient Boosting Model precision on test samples obtained for 50:50 split ratio.

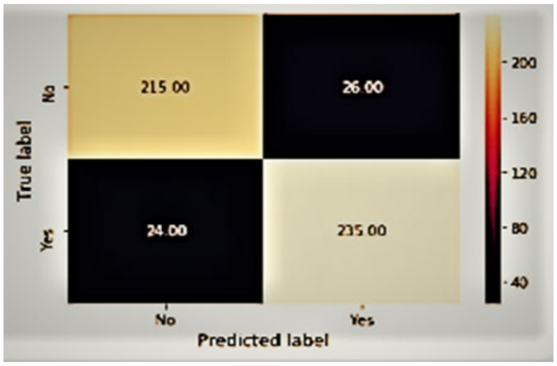


Figure-9. Confusion Matrix for the test samples obtained on 50:50 split ratio for Gradient Boosting Model.

b. XG Boosting :

Here the boosting is applied with 100 weak learners by considering the base estimator as random forest. This technique is considered boosting technique: XG Boosting Algorithm.

Table 8 Predictive Performance of XG Boosting Model.

Sl No	Split Ratio	Accuracy	Precision	Recall	F1-Score	ROC Curve
1	70:30	90.0	91.0	87.0	89.0	0.95
2	60:40	90.0	90.0	88.0	89.0	0.95
3	50:50	**90.0**	**91.0**	**89.0**	**90.0**	**0.95**

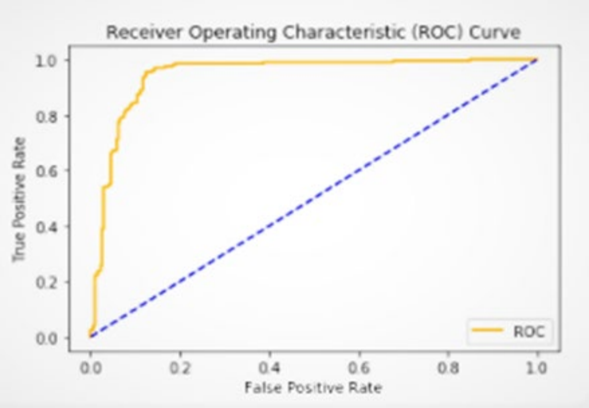


Figure 10. Receiver Operating Characteristic curve for XG Boosting Model precision on test samples obtained for 50:50 split ratio**.**

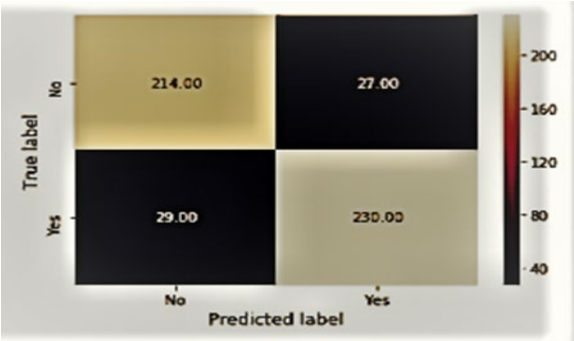


Figure 11 Confusion Matrix for the test samples obtained on 50:50 split ratio for XG Boosting Model.

c. Ada Boosting :

Here the boosting is applied with 100 weak learners by considering the base estimator as random forest. This technique is considered for different boosting technique: Ada Boosting Algorithm.

Table 9 Predictive Performance of Ada Boosting Model.

Sl No	Split Ratio	Accuracy	Precision	Recall	F1-Score	ROC Curve
1	70:30	88.0	88.0	88.0	88.0	0.93
2	60:40	86.0	86.0	87.0	88.0	0.92
3	50:50	**87.0**	**87.0**	**87.0**	**87.0**	**0.93**

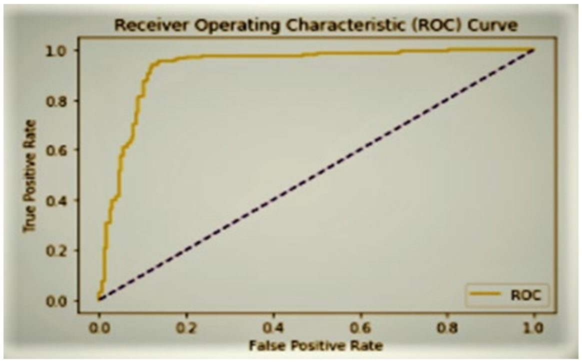


Figure 12 Receiver Operating Characteristic curve for Ada Boosting Model precision on test samples obtained for 50:50 split ratio.

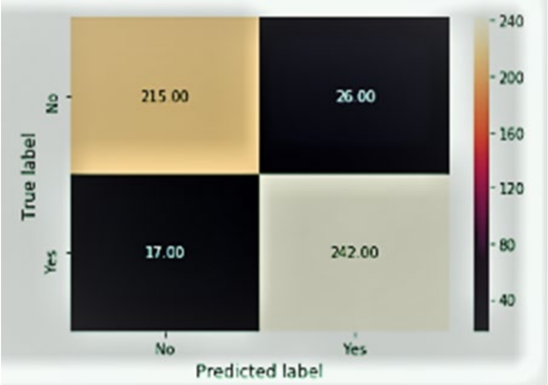


Figure 13 Confusion Matrix for the test samples obtained on 50:50 split ratio for Ada Boosting Model.

The different split performance measures is tabulated in Table 7, Table 8 and Table 9 for Gradient Boosting, XG Boosting and Ada Boosting respectively. For 50:50 split ratio the C= 90%, C=90%, C=87% with Recall of 90%, 89%, 87%, and AUC ROC test of 0.95, 0.95 , 0.93(figure 8, figure 10 and figure 12) have obtained and tabulated for Gradient boosting, XG boosting and Ada boosting on test samples respectively for 100 weak learners by keeping ramdom forest as the base estimator.

The confusion matrix for test samples is shown in figure 9, figure 11 and figure 13 for Gradient boosting, XG boosting and Ada boosting respectively. The algorithms identifies 215,214,215 True positive rate and 235,230,242 True negative rates .Also it has observed 24, 29, 17 of the observations are negative but it predicted positive whereas 26,27,26 observation are postive for the test samples for Gradient boosting, XG boosting and Ada boosting respectively. By looking the overall observations, we can state the Gradient boosting and XG Boosting are similar more accurate when compaired to Ada boosting algorithm for 100 weak learners for random forest as the base estimator.

Out of 7 potential indicators, the decision tree distinguished the accompanying 5 variables as key to yield expectation: Age, OC, P, K, and pH are the key variables. This was steady with the outcomes got utilizing old style measurable techniques on the equivalent dataset. None of the leftover 2 factors (Zone and Yield) were utilized by decision tree to anticipate Yield. Moreover, the node division in the decision tree model give a sign with respect to what explicit levels of the Age,OC, P, K, pH were precisely correlated with every Yield +ve and Yield -ve. For illustration, the Decision Tree identified that all coffee details are correctly identified by grouping them with the expected yield class.

The scatter plot visualized in figure 14 analyzed the variation of yield obtained in different zones considered like: 1- Balehonnur , 2- Belur, 3-Chikmagaluru Town, 4- Hassan, 5- Kalasa, 6- Koppa, 7- Mudigere, 8- N R Pura and 9- Sakleshpura with respect to Age for every 10 intervals between age from1 to 90 and Yield Classes 1 to 4.

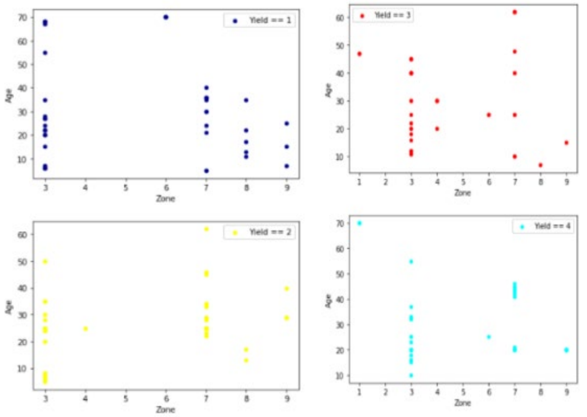


Figure 14 The scatter plot above visualize between Zone and Age with respect to Yield Classes from 1 to 4

4.0 DISCUSSIONS

Six different models are used to predict the Coffee Yield based on the factors associated for the prediction including the continuous and independent variables. Bagging and boosting techniques is also used to build the performance of the model.

Based on the analysis done Random Forest and Extra tree classifier showed the model with more accurate in predicting the Yield with the factors involved like Age, P, pH, K and %OC and less risk factors like Zone.

Extra Tree classifier is the advance version of Decision Tree and helps to reduce the node or feature to increase the presentation of the model. Random Forest is an ensemble method which create multiple classification and regression trees also, look across randomly chosen input factors to decide the split.

The results of our study showed the coffee yield prediction and the most important factors related are functioned in the models used to predict the desired yield based on the data. In other studies, soil parameters were the most important associated factors for predicting the yield. Extra tree classifier is perhaps the most effortless device to decision systems which is built upon decision tree and easy to understand. Projects dependent on these principles can be made and utilized on PCs for decision analysis to conclude the outcomes. In this study, comparison of all models showed that Precision and Recall Values obtained from Random Forest and Extra Tree classifier seems better. On the other hand, Recall value of algorithm was higher than random forest, but precision of random forest and Extra tree classifier was higher than all other models. The reason for being difference between Recall values of them is using different algorithm.

The Receiver Operating Characteristic curve is a technique used to portrait, put together, and pick yield classes in light of the presenting of the classification. However, this technique is used to document about which model performs better and has an undeniable degree of precision. This record, which analyzes the presentation of genuine positive and false positive of two distinctive choice boundaries, is regularly used to assess the prescient precision of characterization models.

In the current study, the AUC consisting of random forest and Extra Tree classifier of testing dataset was significantly higher than decision tree based on Performance measures obtained which is shown in Table-4 and Table-5. Random forest model is an accurate model for investigation of novel predictor markers, which is in line with previous.

The strength of the study lies in its large sample size that makes it applicable to general population. One potential limitation of this study is that, results obtained from longitudinal and cohort data are based on cross-sectional data.

5.0 CONCLUSIONS

1. The objective of this paper is to identify yield class based upon agronomic factors like soil parameters, coffee type, age and zone using machine learning. Hence, we have conducted

AUC ROC calculation on three different machine learning algorithms, in which extra tree classifier (ETC) model provided good prediction due to their efficacy, Precision F1-Score and Recall.

2. And also random forest classifier, boosting algorithms provided good result too and we have developed RF model with 600 bagged Decision Trees (DT's) and 100 weak learners with base estimator as random forest, which furnished powerful arrangement with exact expectation in deciding yield classes for each individual entry in the dataset.

3. This research also includes other machine learning algorithm like Decision tree Classifier which showed lesser accuracy when compared to extra tree classifier random forest classifier and boosting algorithms. This study emphasis for the understanding of various machine learning algorithms which allow us to carry out further analysis using other Predictive models.

Acknowledgment

We would like to thank the Central Coffee Research Institute (CCRI), Balehonnur, Chikkamagaluru District, Karnataka, India for the support through the JSS Science and Technology University (JSSSTU) for providing Historic coffee datasets to carry out our research work towards yield prediction and also for their assistance to understand the various factors involved in coffee which influences to get better yield.

References

[1] Bornemisza, E. 1982. Nitrogen cycling in coffee plantations/Ciclo de nitrógeno en plantaciones de café. *Plant and Soil*,67: 241-246.

[2] Raghuramulu, Y.,Sreenath,H.L. 2016-2017 *Coffee board research department – 70th annual report* , Central coffee research institute, coffee research station, Government of India.

[3] Jaramillo, J., Chabi-Olaye, A., Kamonjo, C., Jaramillo, A., Vega, F. E., Poehling, H. M., & Borgemeister, C. 2009. Thermal tolerance of the coffee berry borer Hypothenemus hampei: predictions of climate change impact on a tropical insect pest. *PloS one*, *4*(8): e6487.

[4] Raghuramulu, Y.,Sreenath,H.L. 2014 *Coffee guide book-a manual of coffee Cultivation*,, Central coffee research Institute (ministry of commerce and Industry, Govt. of india).

[5] Illy, A., & Viani, R. (Eds.). 2005. Espresso coffee: the science of quality. *Academic Press*.

[6] Russell, S. J. 2010. Artificial intelligence a modern approach. *Pearson Education, Inc.*

[7] Alpaydin, E. 2020. *Introduction to machine learning*. MIT press.

[8] Lakshmanan, V., Gilleland, E., McGovern, A., & Tingley, M. 2015. Machine learning and data mining approaches to climate science. In *Proceedings of the 4th International Workshop on Climate Informatics*. 3-246. Basel, Switzerland: Springer International Publishing.

[9] Inza, I., Calvo, B., Armananzas, R., Bengoetxea, E., Larranaga, P., & Lozano, J. A. 2009. Machine learning: an indispensable tool in bioinformatics. In *Bioinformatics methods in clinical research* . 25-48. Totowa, NJ: Humana Press.

[10] Gillison, A. N., Liswanti, N., Budidarsono, S., Van Noordwijk, M., & Tomich, T. P. 2004. Impact of cropping methods on biodiversity in coffee agroecosystems in Sumatra, Indonesia. *Ecology and Society*, *9*(2).

[11] Godsteven P. Maro, Jerome P. Mrema, Balthazar M. Msanya, Bert H. Janssen and James M. Teri. 2014. Developing a Coffee Yield Prediction and Integrated Soil Fertility Management Recommendation Model for Northern Tanzania*. International Journal of Plant and Soil Science* 3(4): 380-396. Article no. IJPSS.005.

[12] Kouadio, L., Deo, R. C., Byrareddy, V., Adamowski, J. F., & Mushtaq, S. 2018. Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. *Computers and electronics in agriculture*, *155*: 324-338.

[13] Romero-Alvarado, Y., Soto-Pinto, L., García-Barrios, L., & Barrera-Gaytán, J. F. 2002. Coffee yields and soil nutrients under the shades of Inga sp. vs. multiple species in Chiapas, Mexico. *Agroforestry systems, 54*: 215-224.

[14] Geurts, P., Ernst, D., & Wehenkel, L. 2006) Extremely randomized trees. *Machine learning*, *63*: 3-42.

[15] Sharma, H., & Kumar, S. 2016. A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, *5*(4): 2094-2097.

[16] Ruß, G., Kruse, R., Schneider, M., & Wagner, P. 2008. Estimation of neural network parameters for wheat yield prediction. In *Artificial Intelligence in Theory and Practice II: IFIP 20 th World Computer Congress, TC 12: IFIP AI 2008 Stream, September 7-10, 2008, Milano, Italy 2*: 109-118. Springer US.

[17] Wang, N., Jassogne, L., van Asten, P. J., Mukasa, D., Wanyama, I., Kagezi, G., & Giller, K. E. 2015. Evaluating coffee yield gaps and important biotic, abiotic, and management factors limiting coffee production in Uganda. *European Journal of Agronomy*, *63*: 1-11.

[18] Bunn, C., Läderach, P., Pérez Jimenez, J. G., Montagnon, C., & Schilling, T. 2015. Multiclass classification of agro-ecological zones for Arabica coffee: an improved understanding of the impacts of climate change. *PLoS One*, *10*(10): e0140490.

[19] Shastry, K. A., Sanjay, H. A., & Deshmukh, A. 2016. A parameter based customized artificial neural network model for crop yield prediction. *Journal of Artificial Intelligence*, *9*(1-3): 23-32.

[20] Sahu, S., Chawla, M., & Khare, N. 2017. An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach. In *2017 international conference on computing, communication and automation (ICCCA)*. 53-57. IEEE.

[21] Mishra, S., Mishra, D., & Santra, G. H. 2016. Applications of machine learning techniques in agricultural crop production: a review paper. *Indian J. Sci. Technol*, *9*(38): 1-14.

[22] Joanes, D. N., & Gill, C. A. 1998. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(1): 183-189.

[23] Schwertman, N. C., Owens, M. A., & Adnan, R. 2004. A simple more general boxplot method for identifying outliers. *Computational statistics & data analysis*, *47*(1): 165-174.

[24] Natekin, A., & Knoll, A. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*: 21.

[25] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 785-794.

[26] Chengsheng, T., Huacheng, L., & Bing, X. 2017. AdaBoost typical Algorithm and its application research. In *MATEC Web of Conferences* 139: 00222. EDP Sciences.